Unlocking Climate Change Scientific Reports: Keyphrase Extraction and Ontology Enrichment

Eirini Katsadaki, Margarita Kokla
School of Rural, Surveying and Geoinformatics Engineering, National Technical University of Athens
9, H. Polytechniou Str., 15780, Zografos Campus, Athens, Greece
eirinikats@mail.ntua.gr; mkokla@survey.ntua.gr

Objective

Climate change reports, such as these prepared by the Intergovernmental Panel on Climate Change (IPCC, 2022) and other organizations contain vital information for comprehending climate change causes, impacts, and interconnections, but the complexity and diverse terminology used makes it challenging to extract and organize relevant information.

Keyphrase extraction is essential for enhancing the understanding and organization of information in these complex scientific texts. It allows for the identification of important concepts and entities, aiding in content visualization, search, retrieval, and question answering. Keyphrase extraction may also support ontology enrichment by forming bridges between natural language and formalized ontologies, improving semantic representation and integration.

In this paper, we implement an approach that leverages automated key phrase extraction from climate change reports by comparing three different approaches, followed by the enrichment of the SWEET (Semantic Web for Earth and Environmental Terminology) Ontology with the highest scoring key phrases.

Methodology

The workflow includes four main steps (Fig. 1):

1. Pre-processing: text cleaning, lemmatization and tokenization.
2. Named Entity Recognition to identify locations and events.
3. Keyphrase Extraction: comparison of Amazon Comprehend, TF-IDF, and Yake.
4. Ontology Enrichment of the SWEET ontology with high-scored keyphrases

Chapter 16 of the Sixth Assessment Report of the Intergovernmental Panel on Climate Change Working Group II entitled "Key Risks across Sectors and Regions" (O'Neill et al., 2022) has been used as per the comparative keyphrase extraction and ontology enrichment process.

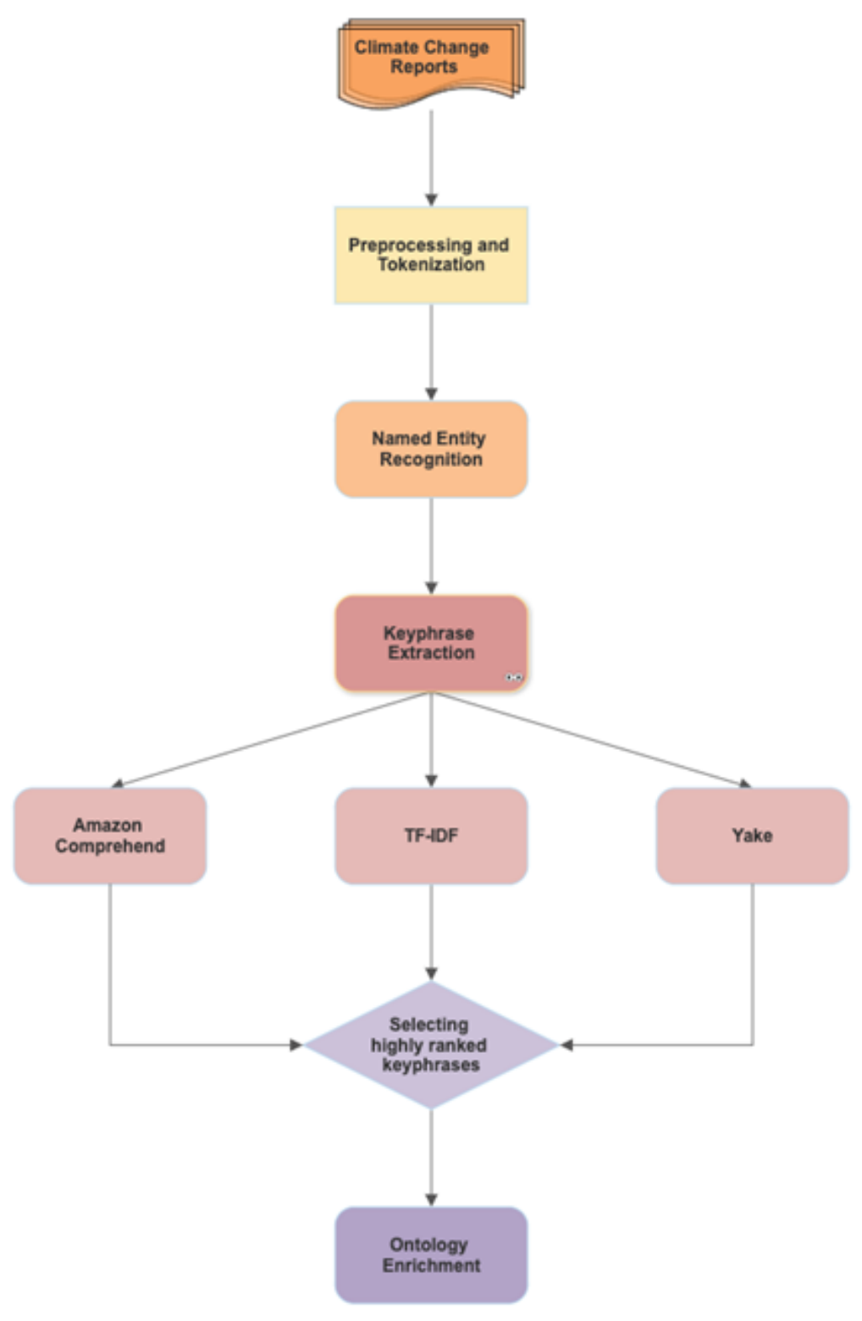


Figure 1. Workflow of the proposed approach

Location extraction and visualization

The second step involved Named Entity Recognition (NER) on the preprocessed text to identify places and events and the subsequent geocoding and visualization of locations using the Python library 'spaCy'(Fig.2).

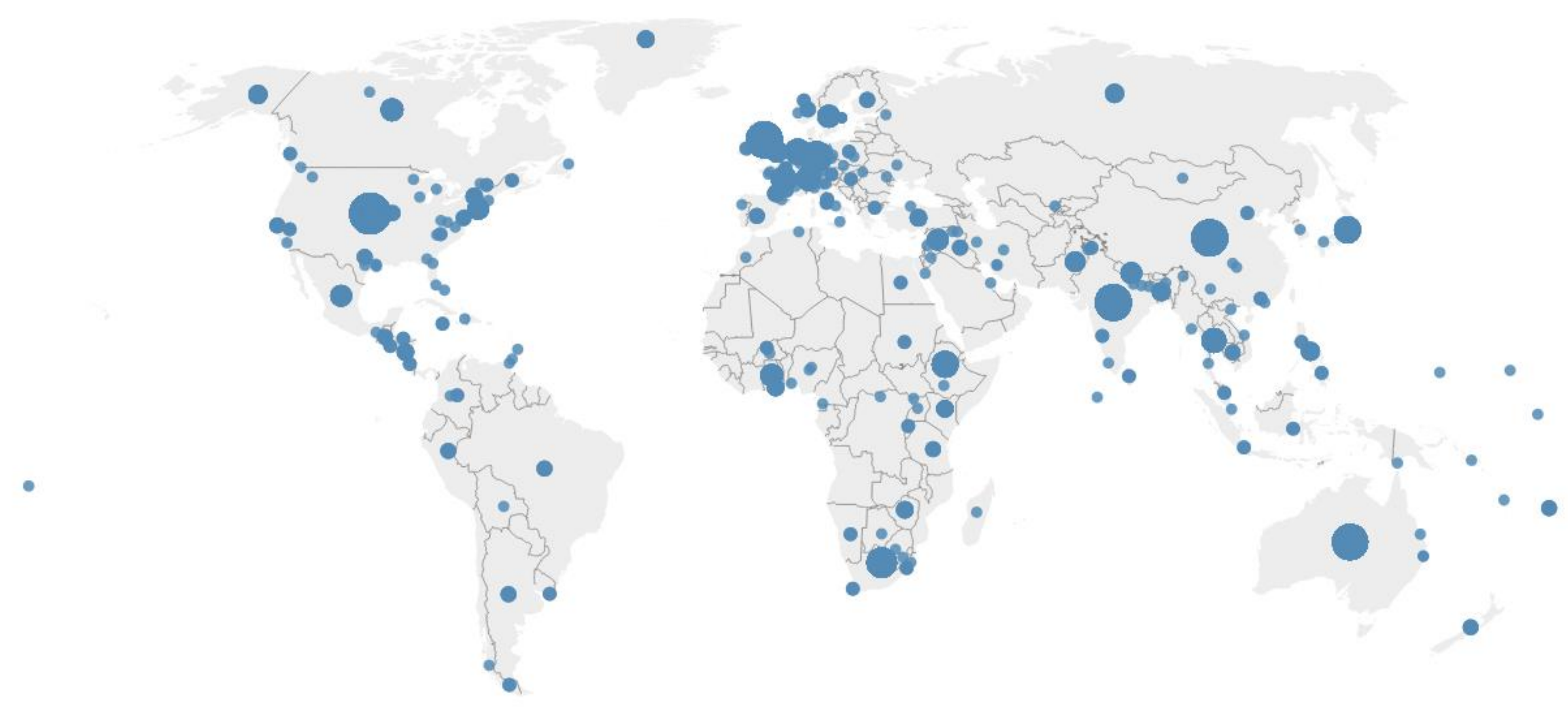
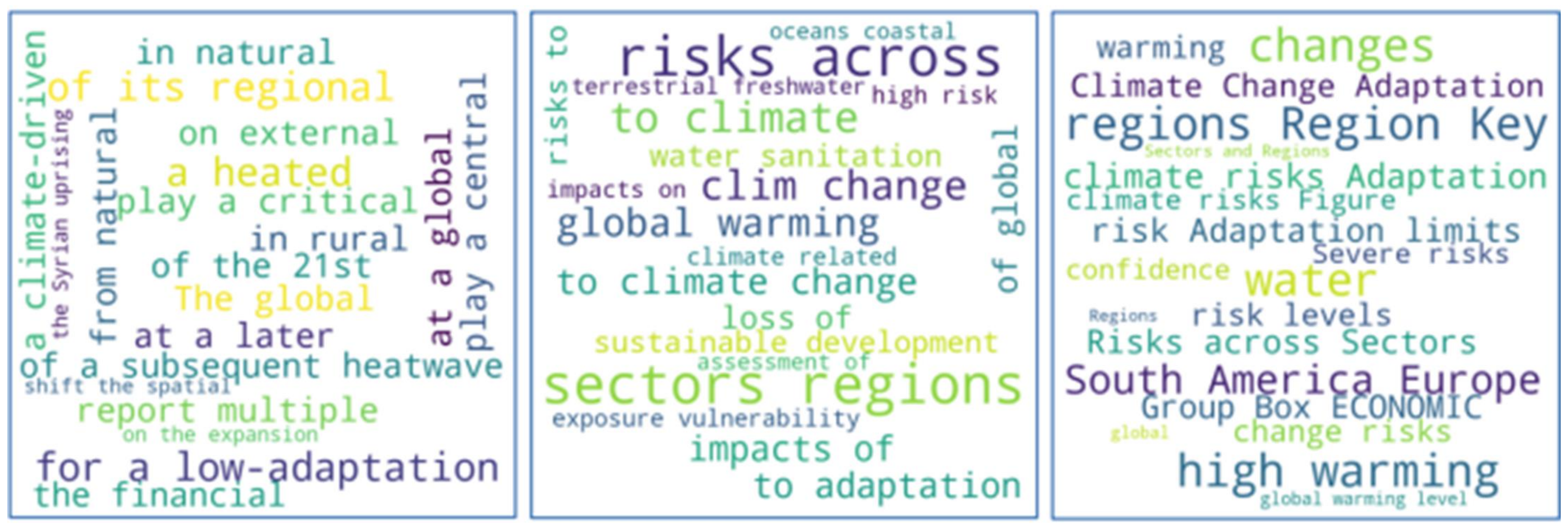


Figure 2: Distribution of extracted place names based on the frequency of reference.

Keyphrase Extraction

The third step employed three distinct approaches for keyphrase extraction to compare their accuracy and effectiveness for the extraction process:

- TF-IDF (Term Frequency-Inverse Document Frequency) is a widely used algorithm for keyphrase extraction that calculates the relevance of a term within a document or corpus by assigning weights to terms based on their frequency within a single document and their inverse frequency across that document (Luhn, 1958) to prioritize terms that appear frequently within the document while being less common overall, highlighting their significance within the context of the document.
- Amazon Comprehend (https://aws.amazon.com/comprehend/) is a web service that uses a combination of statistical techniques, rule-based matching, linguistic heuristics, and deep learning-based models for the extraction of keyphrases.
- YAKE (Yet Another Keyword Extractor) uses a sequence labeling algorithm to identify and extract keyphrases based on their statistical properties, such as their frequency and distribution within the text, as well as their linguistic properties, such as their part of speech and position in the sentence (Campos et al., 2020).



Amazon Comprehend **TF-IDF** **Yake**

Figure 3: Highest Score Concepts

Cosine similarity (Singhal, 2001) was employed to measure the resemblance between keyphrases and the SWEET ontology concepts. It quantifies the cosine of the angle between two vectors, representing the degree of alignment or resemblance between them.

Among the approaches considered, Amazon Comprehend consistently yielded the highest cosine similarity score, followed by TF-IDF, and lastly, Yake (Table 1).

Keyphrase Extraction Approach	Cosine Similarity Score
Amazon Comprehend	34.1%
TF-IDF	22.6%
Yake	1.1%

Table 1: Similarity Score with SWEET Ontology

Ontology Enrichment

The fourth step leveraged the extracted keyphrases with the highest scores for ontology enrichment to create a more comprehensive and extended representation of the specific domain concepts used in the input report.

The SWEET ontology was enriched with:

- Extracted keyphrases which were added as subclasses of SWEET concepts.
- Cause-effect and other association relations between concepts.

New Concepts	Relation	SWEET Concepts
Urban Flood **Coastal Flood** **Fluvial Flood** **River Flood**	SubClass Of	Flood
Agricultural Drought **Ecological Drought**	SubClass Of	Drought
Vector-Borne Disease **Water-Borne Disease** **Food-Borne Disease**	SubClass Of	Disease
Marine Biodiversity **Terrestrial Biodiversity** **Alpine Biodiversity**	SubClass Of	Biodiversity
Human migration	SubClass Of	Migration
Internal migration **International migration** **Urban migration**	SubClass Of	Human migration
Economic impact **Societal impact**	SubClass Of	Impact

Table 2: Examples of keyphrases added as subclasses of SWEET concepts

New Concept	Relation	Sweet Ontology
Climate change risk	Caused-By	Climate change
Internal migration	Linked-To	Extreme event
International migration	Linked-To	Extreme event
Morbidity	Caused-By	Water-borne disease
Mortality	Caused-By	Water-borne disease
Morbidity	Caused-By	Vector-Borne disease
Mortality	Caused-By	Vector-Borne disease
Mortality	Caused-By	Extreme event
Maladaptation	Opposite Of	Adaptation

Table 3: Examples of cause-effect and other associations between concepts

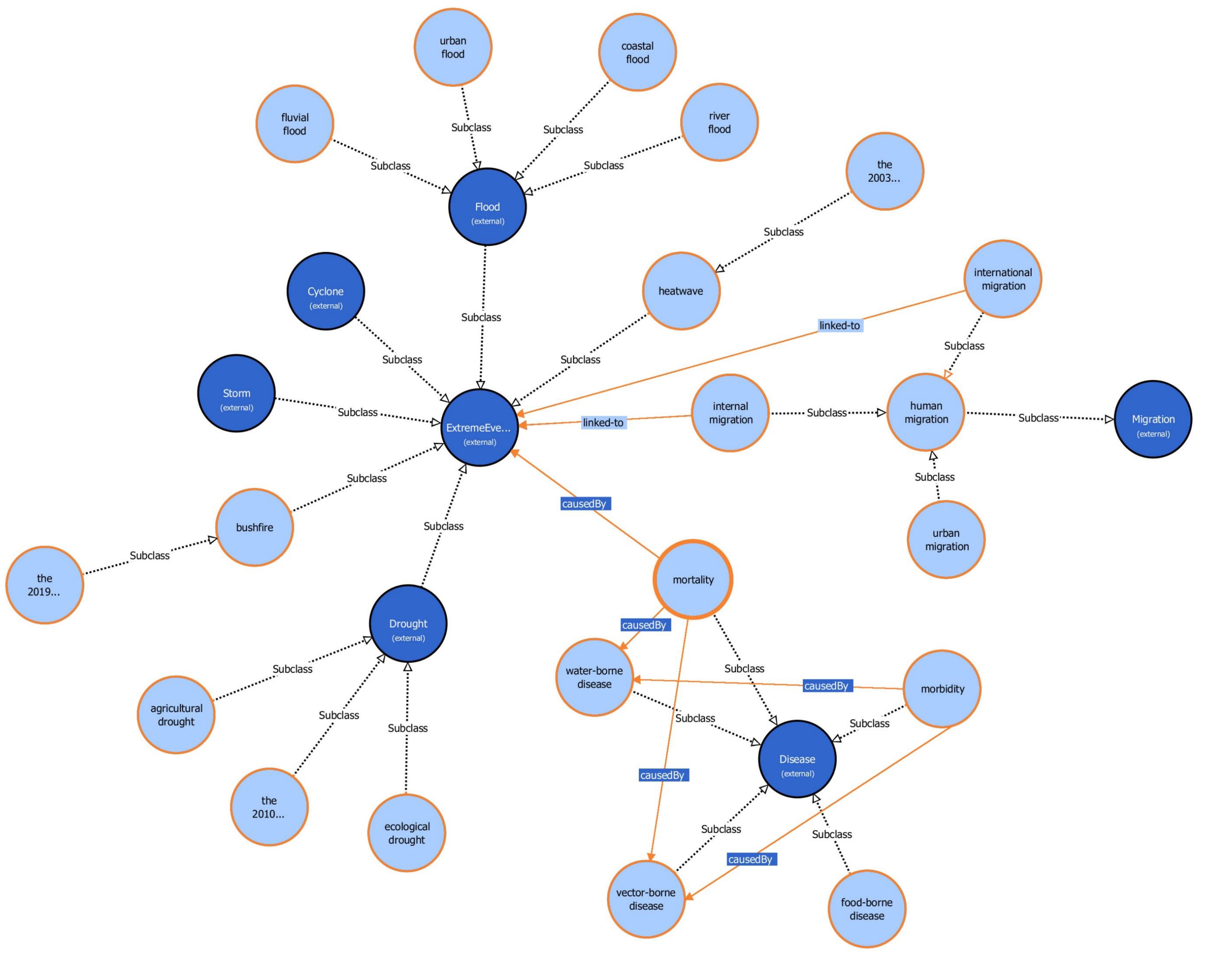


Figure 4: An excerpt of the enriched ontology. The new concepts are shown with orange outlines and the new relations with orange lines.

Conclusions

The resulting enriched ontology captures meaningful relations between concepts, uncovering connections between climate change and factors such as urbanization, poverty, human moblility, maladaptation and other social, economic, and environmental aspects. Moreover, the inclusion of keyphrases related to specific natural disasters, such as droughts, heatwaves, and wildfires has expanded the scope of the ontology and improved its comprehensiveness in capturing complex interactions between climate change and its impacts across regions.

Climate change is a multi-faceted topic, and relation extraction techniques could be used as an additional process to identify the complicated relations between climate change concepts, as well as their connections to specific places on Earth.

References

- Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, *509*, 257-289.
- IPCC, 2022: *Climate Change 2022: Impacts, Adaptation, and Vulnerability.* Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Löschke, V. Möller, A. Okem, B. Rama (eds.)]. Cambridge University Press. Cambridge University Press, Cambridge, UK and New York, NY, USA, 3056 pp., doi:10.1017/9781009325844.
- Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, *2*(2), 159-165.
- O'Neill, B., M. van Aalst, Z. Zaiton Ibrahim, L. Berrang Ford, S. Bhadwal, H. Buhaug, D. Diaz, K. Frieler, M. Garschagen, A. Magnan, G. Midgley, A. Mirzabaev, A. Thomas, and R. Warren, 2022: Key Risks Across Sectors and Regions. In: Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Löschke, V. Möller, A. Okem, B. Rama (eds.)]. Cambridge University Press, pp. 2411–2538, doi:10.1017/9781009325844.025.
- Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, *24*(4), 35-43.

The research project is supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment" (Project Number: HFRI-FM17-2661).